P2: Q&A

1. What was the project objective, and what steps did you take to achieve it?
Our main objectives were to pinpoint key factors affecting salary levels in data-related roles and to build a predictive model using 2024 salary data. To achieve this, we processed the dataset through data wrangling, performed exploratory data analysis (EDA), and developed models using techniques like Ridge, Lasso, Random Forest, and Gradient Boosting Machines.

2. What dataset did you use for this project?
Our team used the Global AI, ML, and Data Science Salary Index from ai-jobs.net. The extracted index provided information on global salary distribution, focusing on data-related jobs. To add more context, the columns included in our datasets were work year, job title, salary, employment type, and company size.

3. How did you wrangle your dataset to ensure it was clean and structured properly?
For our project, we performed several data wrangling tasks. These tasks include correcting data types, handling missing values, addressing data entry errors, standardizing categorical values, and removing/filtering unnecessary columns and rows not needed for the analysis and the model training process.

3. What methods did you use for exploratory data analysis (EDA)?
For EDA, we used histograms to analyze the distribution of our target variable, bar charts to compare the average salary across different categorical features, and box plots to identify outliers. We also built correlation matrices and feature importance bar charts to understand relationships between variables.

4. What insights did you extract from your Salary histogram?
The salary histogram displayed a normal distribution (close to), suggesting that linear regression models could be effective due to their compatibility with normally distributed target variables. Therefore, we prioritized ridge and lasso regression models for our training process.

5. What insights did you extract from your bar charts comparing the average salary with each input feature?
The bar charts displayed a balanced progression of salary increments across different categories. This indicates that there are no dominant categories that could bias the model.

6. How did you handle outliers in your dataset?
The box plot displayed outliers in the salary data beyond the $310,000 mark. As a team, we concluded that these outliers were legitimate and were not removed for model training to reflect actual high-salary cases. In one iteration, we experimented by capping these outliers at $310,000 to assess their impact on the model’s performance.

7. Which machine learning models did you consider, and why?
We chose four models for our project (Ridge, Lasso, Random Forest, and Gradient Boosting Machines). As mentioned in the previous question, Ridge and Lasso were selected as our target variable, salary, displayed a normal distribution (referenced from the histogram). Also, both linear regression models apply regularization techniques to prevent overfitting and improve model performance. Random forest and Gradient Boosting Machines were selected to experiment with non-linear regression models and assess how they'll perform with our dataset.

8. Could you summarize your model training process?
We performed three iterations of model training for each selected model. The first iteration included all features. The second iteration addressed salary outliers by capping extreme values. The third iteration used the top three features based on the feature importance bar chart plotted for each model. After 12 iterations, the best-performing model was selected as our final selection. Only to this final model, we adjusted the hyperparameters using grid search to optimize the model's performance.

9. Where were the results of your model training process?
The Gradient Boosting Machines model displayed the best R2 score out of all the models. However, the final model did not perform as well as expected. We suspect that the reason for the low performance is due to the dataset lacking essential information for accurate salary prediction, like economic indicators. We also recognize the need to improve our feature engineering to optimize our encoding process so that our models can learn from the data properly.

10. What are the next steps for the model?
For us, the next steps are very straightforward. We would like to add more economic indicators to the dataset, improve our feature engineering and model selection strategies, and apply cross-validation techniques.